Cristina Vicarelli
Avvocato

I big data sono entrati nella statistica

Nella newsletter del Garante Privacy pubblicata in data 16 ottobre 2014, si legge che i **Big Data** sono entrati nelle statistiche nazionali, grazie al parere favorevole reso dall'Autorità in ordine al Programma statistico nazionale 2014-2016 Aggiornamento 2015-2016 (Psn), predisposto dall'Istat.

Stando alla comunicazione del Garante, il Piano prevede, tra l'altro, la possibilità di utilizzare per la prima volta a fini statistici, seppur in via sperimentale, i Big Data di telefonia mobile. Tale elaborazione statistica avrebbe l'obiettivo di effettuare una stima a livello aggregato dei flussi di mobilità intercomunali delle persone, utile per la programmazione e la gestione dei servizi locali e l'individuazione di opportune misure di Protezione civile.

Il Garante tuttavia sintetizza nella newsletter e nel parere una materia complessa, e non è facile trarre indicazioni per applicazioni che trascendano il caso concreto e per chiarire ai più, al di là del dato lessicale, come siano intesi i big data dall'Autorità e quale impiego se ne possa fare nel nostro ordinamento.

Dati anonimizzati e dati pseudonimizzati

Il primo problema che balza agli occhi concerne l'anonimizzazione dei big data. Si tratta di un'operazione tutt'altro che scontata, tanto che più voci autorevoli hanno paventato la fine dell'anonimato nell'era dei big data, attesa l'incompatibilità tra questa tipologia di elaborazione (che in sintesi estrema e per quello che qui interessa, presuppone l'acquisizione di un gran numero di dati riferibili ad un unico soggetto), e il trattamento in forma anonima che postula l'impossibilità di riassociare i dati ad un soggetto identificabile. Ciò implica un largo ricorso a tecniche di **pseudonimizzazione**, tanto che quest'anno il Gruppo di Lavoro ex art. 29 ha sentito l'esigenza di intervenire sul tema, con un parere strutturato, reso sulle tecniche di anonimizzazione, volto a delineare il distinguo tra dato

anonimo e dato pseudonimizzato (Parere 05/2014 sulle tecniche di anonimizzazione – adottato il 10 aprile 2014).

Per iniziare la mia analisi, prenderò a prestito le definizioni sintetizzate nel **Manuale sul diritto europeo in materia di protezione dei dati** pubblicato dal FRA, "I dati sono *anonimizzati* **se tutti gli elementi identificativi sono stati eliminati** da un insieme di dati personali. Le informazioni non devono mantenere alcun elemento identificativo che, con un *ragionevole sforzo*, potrebbe servire a identificare nuovamente la persona o le persone interessate. Una volta resi completamente anonimi, i dati non sono più ritenuti personali".

Per i dati *pseudonimizzati*, si legge (pag. 46 e ss.) "Le informazioni personali contengono elementi identificativi come nome, data di nascita, sesso e indirizzo. Quando le informazioni personali vengono pseudonimizzate, **gli elementi identificativi sono sostituiti da uno pseudonimo**, che si ottiene, per esempio, crittografando gli elementi identificativi contenuti nei dati personali".

Il progetto dell'ISTAT

Il progetto sottoposto al vaglio della nostra Autorità Garante, prevedeva che l'Istat trattasse dati relativi al cosiddetto **"call detail record" (cdr)**. Il cdr è un **numero progressivo**, assegnato dal gestore telefonico all'utente che effettua la chiamata (in sostituzione del *codice fiscale, nome e cognome*), al quale vanno aggiunte le informazioni relative al *Comune* nel quale si trova la cella di effettuazione, la *data* e l'*ora* della chiamata. Si tratta pertanto, a mio parere, di una tecnica di pseudonimizzazione che come vedremo, però, non esaurisce il trattamento. Infatti si legge nel documento dell'Autorità: "**Gli utenti verranno distinti in quattro categorie**: residenti stanziali, temporaneamente dimoranti, pendolari giornalieri e visitatori occasionali. Riguardo all'uso di questi dati, su specifica richiesta del Garante, l'Istat ha fornito idonee garanzie sul fatto che presso il gestore siano raccolti solo dati in forma anonima. E' stato previsto infatti che il gestore assegni un codice ad ogni cdr e che successivamente venga **eliminata ogni possibilità di raccordo** tra tale codice e gli identificativi originali".

E' comprensibile che il Garante non si soffermi su tali idonee garanzie, che l'istituto di statistica ha, probabilmente, interesse a mantenere riservate.

Tuttavia se il parere si legge in coordinazione con le indicazioni del Gruppo di Lavoro ex art. 29, fornite nel parere sopra richiamato, le tecniche di anonimizzazione appaiono oltremodo oscure.

Parere 5/2014 del Gruppo di Lavoro ex art. 29

In particolare, secondo il Gruppo di Lavoro ex art 29, è essenziale comprendere che quando un **titolare del trattamento non cancella i dati originali (identificabili)** a livello di evento, **e trasmette poi parte di questo insieme di dati** (ad esempio, dopo l'eliminazione o il mascheramento dei dati identificabili), **l'insieme di dati risultante contiene ancora dati personali**. Soltanto se il titolare del trattamento **aggrega** i dati a un livello in cui i singoli eventi non sono più identificabili si può definire anonimo l'insieme di dati risultante. Ad esempio, se un'organizzazione raccoglie dati sugli spostamenti delle persone, i tipi di spostamenti individuali a livello di evento rientrano ancora tra i dati personali per tutte le parti coinvolte, fintantoché il titolare del trattamento (o altri) ha ancora accesso ai dati non trattati originali, anche se gli identificatori diretti sono stati espunti dall'insieme dei dati forniti a terzi. Tuttavia, se il titolare del trattamento cancella i dati non trattati e fornisce a terzi solamente statistiche aggregate ad alto livello, ad esempio "il lunedì sulla rotta X i passeggeri sono più numerosi del 160% rispetto al martedì", i dati possono essere definiti anonimi.

Sempre secondo il Gruppo di Lavoro, un'efficace soluzione di anonimizzazione **impedisce a tutte le parti di identificare una persona in un insieme di dati**, di collegare due dati all'interno di un insieme di dati (o tra due insiemi distinti di dati) e di dedurre informazioni da tale insieme di dati. In generale, eliminare elementi direttamente identificanti non è pertanto di per sé sufficiente a garantire che l'identificazione della persona interessata non sia più possibile. Spesso è necessario adottare misure supplementari per prevenire l'identificazione, ancora una volta a seconda del contesto e degli scopi del trattamento cui sono destinati i dati resi anonimi.

(cfr PAG 10 Parere 5/2014).

Ed ancora: **uno degli errori frequenti**, secondo il Gruppo di Lavoro, **consiste nel ritenere che un insieme di dati pseudonimizzati sia anonimizzato**: spesso i responsabili del trattamento presumono che eliminare o sostituire uno o più attributi sia sufficiente per

rendere anonimo un insieme di dati. Molti esempi hanno dimostrato l'erroneità di tale convinzione; la semplice modifica dell'identità non impedisce l'identificazione di una persona interessata se l'insieme di dati continua a contenere quasi-identificatori o se i valori di altri attributi consentono comunque di identificare una persona. In molti casi identificare una persona all'interno di un insieme di dati pseudonimizzato può essere facile come con i dati originali. Occorre adottare misure supplementari per poter considerare l'insieme di dati effettivamente anonimizzato, tra cui l'eliminazione e la generalizzazione degli attributi o la cancellazione dei dati originali o almeno la loro estrema aggregazione.

(pag 23 Parere 5/2014).

La soluzione adottata dal Garante

Nel parere del Garante si legge che "è stato previsto infatti che il gestore **assegni un codice ad ogni cdr e che successivamente venga eliminata ogni possibilità di raccordo** tra tale codice e gli identificativi originali", *dal che si evince che gli originali non vengano cancellati.*

Appare quindi difficile, nell'estrema sintesi con la quale si riferiscono le operazioni nel parere, definire i canoni che consentono di individuare questi dati come anonimi piuttosto che come pseudonimi.

Differenza molto importante, dato che, come si è visto più sopra, i dati anonimi non sono più dati personali, mentre i dati pseudonimizzati lo sono ancora, con tutte le ricadute normative in tema di consensi e informative del caso.

Si evince che i dati saranno acquisiti dall'Istat in forma aggregata: si legge infatti nel parere: "Al riguardo, l'Istituto ha, preliminarmente, puntualizzato che l'obiettivo della sperimentazione è quello di stimare i flussi intercomunali a livello aggregato e non individuale. Con riferimento al rischio di identificazione derivante dalla possibilità di individuare un'unità elementare tramite collegamento – indiretto – ad altre fonti di dati in possesso dell'Istituto, su specifica richiesta dell'Ufficio del Garante è stata, inoltre, fornita idonea assicurazione che, nell'ipotesi in cui dovessero verificarsi frequenze di flusso inferiori a tre unità, le stesse verranno oscurate".

Ebbene, la compatibilità tra finalità statistiche e big data (e quindi le possibilità e i limiti dell'impiego di questi nell'ambito delle esenzioni previste per i trattamenti a "scopo statistico") è stata indagata nel parere **WP ART 29 n. 3/2013 "on purpose limitation"** (mai tradotto nella nostra lingua) e avrebbe meritato qualche parola in più, essendo tutt'altro che scontata, come si assume nel parere del Garante. A tal proposito si rileva come buona parte del trattamento venga svolta ad opera dell'operatore di telefonia, che non agisce per finalità proprie, ma dell'Istat: circoscrivere l'ambito di esenzione dello scopo statistico sarebbe stato più che opportuno. L'Autorità nazionale, inoltre, pare preoccuparsi delle possibilità di collegamento con altre fonti di dati all'interno dell'Istituto, ma, come si è visto, per il Gruppo di Lavoro è indifferente che le fonti si trovino presso il titolare che attualmente tratta i dati, possono anche trovarsi nella disponibilità di chi glieli ha comunicati.

Anche sotto questo profilo si sorvola su aspetti che sarebbe stato bene indagare.

Il parere e la newsletter appaiono poco puntuali nel delineare scopi, ruoli e legittimazione dei trattamenti dei dati telefonici, e davvero poco esaustivi del definire le procedure di anonimizzazione, rischiando di ingenerare, nel lettore, confusione con la pseudonimizzazione.

Si auspica che l'Autorità voglia tornare sull'argomento presto, in maniera più diffusa, dato anche l'interesse che la Commissione UE ha manifestato proprio in questi giorni rispetto ai big data e attesa la sempre maggiore importanza strategica che paiono destinati ad acquisire anche nel Vecchio Continente, ove il parametro del corretto trattamento dei dati personali, a differenza di quanto avviene oltreoceano, non è certo di secondo momento, e ove sono proprio gli aspetti legati alla privacy che costituiscono il maggior freno al loro sviluppo.